Accred Qual Assur (2007) 12:365–368
DOI 10.1007/s00769-007-0275-4

DISCUSSION FORUM

A new approach in assessing microbiological results in water analysis proficiency testing

H. R. Veenendaal · P. M. van Berkel · G. de Jong · P. K. Baggelaar

Received: 30 November 2006 / Accepted: 19 March 2007 / Published online: 20 April 2007

Abstract Due to the relatively large spread in the results of microbiological proficiency tests, the *z*-scores are often not able to detect zero or low results as being 'bad' results. This paper describes an adapted *z*-score based on the average or the standard deviation of the 50% 'highest' results. The combination of the adapted *z*-scores of four samples enables a better judgement of the performance of each laboratory.

Keywords Proficiency testing · Microbiology · *z*-scores

Introduction

Since the early 1970s, Kiwa Water Research has organised organic, inorganic and microbiological proficiency tests in various water matrices. The primary objective of the Kiwa proficiency tests is to create an opportunity for laboratories to test their own performance under conditions that represent daily practice. This means that participants receive samples made of practice water matrices and apply their own methods. For almost 10 years, the Kiwa Water Research Proficiency Testing Services organisation has been accredited by the Dutch Council for Accreditation. This accreditation guarantees participants and other stakeholders high-quality samples (e.g. homogeneous, stable, compatible with matrices offered in practice), suitable statistics, clear reports and impartiality of the organiser.

The Kiwa Water Research proficiency tests annually consist of approximately 30 laboratory test comparisons in different types of water, i.e. drinking water, surface water, waste water, ground water and swimming water. More than 100 inorganic and organic parameters and microbiological organisms are involved. The statistical processing of results of the chemical proficiency tests is based on Youden statistics. With these statistics, it is possible to establish if deviating results are caused by systematic errors and/or relatively large random errors.

For assessment of the individual performance of a laboratory, *z*-scores are used [1].

Assessment and evaluation of microbiological results

In the Kiwa Water Research proficiency tests, for each organism, the performance of an individual laboratory is assessed by calculating a *z*-score based on group average results using the following formula:

$$Z_i = \frac{x_i - \bar{x}}{s}$$

with x_i being the result of laboratory i, $\bar{x}$ being the average result of all participating laboratories (group average result) and s being the standard deviation of these results.

For the assessment, the following criteria are used:

– A *good* performance with regard to the group average when $|Z_i| \leqslant 2.0$

Papers published in this section do not necessarily reflect the opinion of the Editors, the Editorial Board or the Publisher.

H. R. Veenendaal (✉) · G. de Jong · P. K. Baggelaar
Kiwa Water Research, P.O. Box 1072,
Nieuwegein 3430 BB, The Netherlands
e-mail: harm.veenendaal@kiwa.nl

P. M. van Berkel
Netherlands Forensic Institute,
The Hague, The Netherlands

– A *moderate* performance with regard to the group average when $2.0 < |Z_i| \leqslant 3.0$
– A *bad* performance with regard to the group average when $|Z_i| > 3.0$

The evaluation of our microbiological proficiency test results of recent years showed that, with this *z*-score, very often, it was not possible to distinguish between good and bad results; this was mainly due to the large spread in the participants' results. According to the *z*-score, most laboratories performed *good*, although one would expect, based on the reported results, that a part of these laboratories performed *moderate* or even *bad*. It even incidentally occurred that laboratories who reported zero for samples which actually contained rather high concentrations of the target bacteria were still assessed as having a *good* performance.

This problem would be solved if the *z*-score could be calculated using the average and standard deviation of the reference sample. But these values cannot be obtained with enough precision.

Adapted *z*-score and overall judgement

To solve the above-mentioned problem, two new approaches were introduced, i.e. an *adapted* *z*-score for each individual sample and an *overall judgement* for a laboratory, based on its adapted *z*-scores for all of the samples in the proficiency test.

The *adapted* *z*-score is based on the average and the standard deviation calculated from 50% of the 'highest' results (all results above the median) after performing the Grubbs' test on outliers [2]. In the following formula:

$$Z_i^* = \frac{x_i - \bar{x}^*}{s^*}$$

in which x_i is the result of laboratory i (i=1,..., n) and $\bar{x}^*$ and s^*, respectively, are the average and the standard deviation as calculated from the n/2 highest values. The assessment criteria are the same as for the *standard* *z*-score.

For the organisms *E. coli*, bacteria of the Coli-group, enterococci, *Clostridium perfringens*, *Aeromonas*, *Legionella*, *Pseudomonas aeruginosa* and Staphylococci, a 'high' result can be seen as an indication that the laboratory is capable of performing the analysis. A 'high' result will seldom be caused by introduction of the organism by the participant. It is unlikely that the *adapted* *z*-score will be influenced by very high values caused by dilution errors or the use of false volumes, as these will be filtered by Grubbs' test on outliers. However, for parameters like plate count at 22°C and 36°C, colonies on R_2A and ATP, it is very likely that a 'high' result is caused by a contamination introduced by the participant. Therefore, for these parameters, the *standard* *z*-score based on the average group results is applied.

Our experience with microbiological proficiency tests is that, in most of the test samples, the distribution of the results is bimodal, often with two rather similar peaks (Fig. 1). This justifies the choice of the 50% highest results for the calculation of the adapted *z*-score.

In the Kiwa Water Research microbiological proficiency tests, participants receive four different samples. Each of these samples is judged [*good* (G), *moderate* (M) or *bad* (B)], by either using the *standard* or *adapted* *z*-score, depending on the type of organism. These four individual judgements are then recombined to an overall assessment of the laboratory performance, the so-called *overall judgement*.

Monte-Carlo simulation

With Monte-Carlo simulation, the chances were determined for scoring a *good*, *moderate* or *bad* performance for one sample, using the 50% 'highest' values in a random sample of n values from a normal distribution. We also determined the chances for the combinations of these judgements in four samples.

This Monte-Carlo simulation was applied with the following steps:

1. Randomly draw n values from a normal distribution with an average of 100 and a standard deviation of 5
2. Sort these values from low to high and take the n/2 highest values
3. Estimate the average and the standard deviation of the population of these n/2 highest values
4. Calculate for each of the n values the *adapted* *z*-score (Z_i^*)
5. Count the number of *z*-scores in the following five intervals: $Z^* < -3$; $-3 \leq Z^* \leq -2$; $-2 \leq Z^* \leq +2$; $+2 \leq Z^* \leq +3$; $+3 < Z^*$
6. Repeat steps 1 to 5 100,000 times
7. Determine for each interval the percentage of counts relative to the total number of *z*-scores (100,000×n)
8. Repeat this simulation for n=10, 20, 40, 60 and 100 (n can be seen as the number of participants in a proficiency test)

Results

The results of the above-mentioned simulation are given in Table 1.

A Grubbs' outlier will always receive the judgement 'bad.' When these chances are added for each judgement,

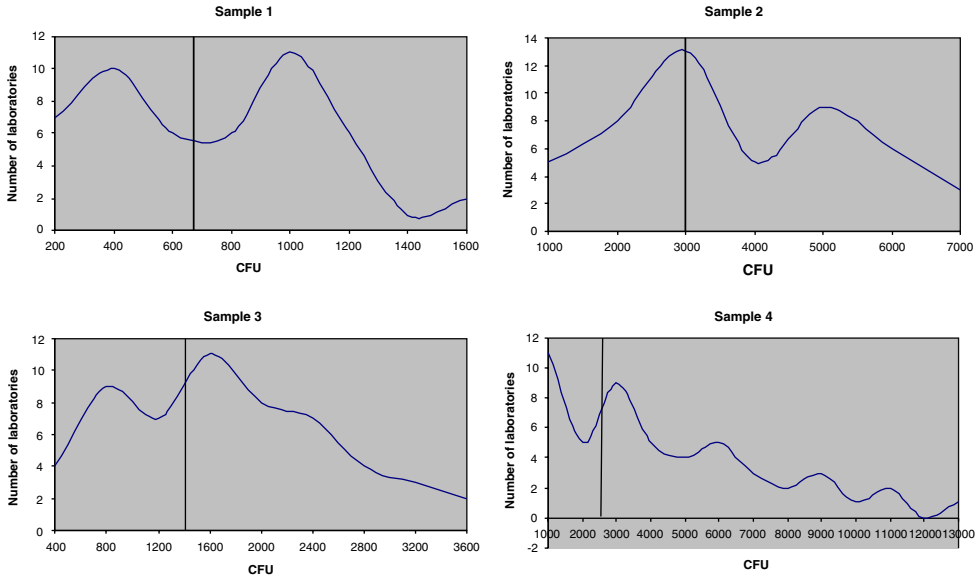

Fig. 1 Distribution of results of four samples in a proficiency test. Results after Grubbs' test on outliers. The *vertical line* indicates the median of the results

Table 1 Results from the Monte-Carlo simulation

z-score	Judgement	n=10 (%)	n=20 (%)	n=40 (%)	n=60 (%)	n=100 (%)
$Z^* < -3$	Bad	17.52	16.72	16.23	16.02	15.88
$-3 \leq Z^* < -2$	Moderate	13.67	15.91	17.12	17.60	17.96
$-2 \leq Z^* \leq +2$	Good	68.81	65.81	64.61	64.24	63.97
$+2 < Z^* \leq +3$	Moderate	0.00	1.57	1.87	1.87	1.84
$+3 < Z^*$	Bad	0.00	0.00	0.17	0.27	0.35
Total		100.00	100.00	100.00	100.00	100.00

Z^*=adapted z-score

the following chances for the various judgements for one sample (Table 2) are obtained.

The chances of the various possible combinations of the individual judgements in the case of four samples are given in Table 3. The number of permutations reflects the number of different sequences in which the given combination can be present in the four samples. The last column of Table 3 shows which *overall judgement* is best suited for the four individual judgements.

From Table 3, it is clear that the chance of obtaining four 'bad' judgements is negligible (<0.1%) if the results from all laboratories come from the same normal distribution (all of the laboratories have the same accuracy). In other words, when a laboratory obtains four *bad* judgements, this may be seen as a clear indication that it performs worse than the others in the proficiency test.

The total chances for each of the three *overall judgements* per number of participating laboratories in a proficiency test is given in Table 4.

So, when all laboratories have the same accuracy, the chance is approximately only 0.4% that a laboratory will get the overall judgement 'bad.'

Discussion

Using the values of a reference laboratory for microbiological proficiency tests is troublesome because the judgement of the performance of the participants will highly depend on the performance of that single reference laboratory. We propose to use an adapted z-score based upon the assumption that the highest results (while using

Table 2 Chances for each possible judgement in the case of one sample

One sample	n=10 (%)	n=20 (%)	n=40 (%)	n=60 (%)	n=100 (%)
Chance 'good'	68.81	65.81	64.61	64.24	63.97
Chance 'moderate'	13.67	17.48	18.99	19.47	19.80
Chance 'bad'	17.52	16.72	16.40	16.29	16.23

Table 3 Chances of the various possible combinations of the individual judgements in the case of four samples and the recommended overall judgement

Chance for four samples	Permutations	*n*=10 (%)	*n*=20 (%)	*n*=40 (%)	*n*=60 (%)	*n*=100 (%)	Judgement
[GGGG]	1	22.41	18.75	17.43	17.03	16.74	Good
[GGGM]	4	17.82	19.92	20.49	20.65	20.73	Good
[GGGB]	4	22.83	19.05	17.69	17.28	16.99	Good
[GGMM]	6	5.31	7.94	9.03	9.39	9.63	Good
[GGMB]	12	13.61	15.18	15.60	15.71	15.78	Good
[GGBB]	6	8.72	7.26	6.73	6.57	6.47	Good
[GMMM]	4	0.70	1.41	1.77	1.90	1.99	Good
[GMMB]	12	2.70	4.03	4.58	4.76	4.89	Good
[GMBB]	12	3.47	3.86	3.96	3.98	4.00	Moderate
[MMMM]	1	0.03	0.09	0.13	0.14	0.15	Moderate
[GBBB]	4	1.48	1.23	1.14	1.11	1.09	Moderate
[MMMB]	4	0.18	0.36	0.45	0.48	0.50	Moderate
[MMBB]	6	0.34	0.51	0.58	0.60	0.62	Moderate
[MBBB]	4	0.29	0.33	0.33	0.34	0.34	Bad
[BBBB]	1	0.09	0.08	0.07	0.07	0.07	Bad
Total	81	100.00	100.00	100.00	100.00	100.00	

Table 4 Total chances for each of the overall judgements

Overall judgement	*n*=10 (%)	*n*=20 (%)	*n*=40 (%)	*n*=60 (%)	*n*=100 (%)
Good	94.1	93.5	93.3	93.3	93.2
Moderate	5.5	6.0	6.3	6.3	6.4
Bad	0.4	0.4	0.4	0.4	0.4

selective culture media) are the best results. For most microbiological parameters, low results will be caused by moderate or bad culture media, wrong incubation temperatures or improper confirmations. It is not likely to find more target organisms in a sample than were added, since the introduction of these target organisms during analysis is highly unlikely. Results that are too high can be caused by using wrong volumes or by improper confirmations. However, these results will mostly be eliminated by the Grubbs' test on outliers. In most of the Kiwa Water Research microbiological proficiency tests, the results show a bimodal distribution, where the peak of the highest results approximately makes up half the results, sometimes some more, sometimes some less. Therefore, it is proposed to use the average and standard deviation of the 50% highest values (after performing the Grubbs' test on outliers) for the calculation of the *adapted z*-score. These are all of the results above the median. By then recombining the individual judgements of the four results of a laboratory in a proficiency test, a more realistic assessment of its performance can be made.

Conclusions

The *adaptedz*-score approach (based on the 50% 'highest' results), in combination with an *overall judgement* for all four samples in the test will enable a more realistic assessment of the performance of a laboratory in microbiological proficiency tests.

References

1. Thompson M, Wood R (1993) The international harmonized protocol for the proficiency testing of (chemical) analytical laboratories. J Pure Appl Chem 65(9):2123–2214
2. Grubbs FE (1969) Procedures for detecting outlying observations in samples. Technometrics 11(1):1–21